ED 374 161 TM 022 090

AUTHOR Wheeler, Patricia H.
TITLE Methods for Assessing Performance. EREAPA Publication Series No. 93-6.
INSTITUTION EREAPA Associates, Livermore, CA.
PUB DATE 93
NOTE 11p.
AVAILABLE FROM EREAPA Associates, 2840 Waverly Way, Livermore, CA 94550-1740 ($4).
PUB TYPE Reports - [illegible]tive/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Adaptive Testing; Assessment Centers (Personnel); Audiotape Recordings; Check Lists; Computer Assisted Testing; *Criteria; Critical Incidents Method; *Educational Assessment; Essays; *Evaluation Methods; Interviews; Journal Writing; *Measurement Techniques; Observation; Portfolios (Background Materials); Rating Scales; Simulation; Student Projects; Teaching Methods; *Test Use; Videotape Recordings; Work Sample Tests
IDENTIFIERS Anecdotal Records; In Basket Simulation; Paper and Pencil Tests; *Performance Based Evaluation

ABSTRACT

A variety of assessment methods and instruments can and should be used to evaluate the performance of individuals and groups. Some possible assessments are described, including (1) adaptive testing; (2) anecdotal records; (3) assessment centers; (4) behavioral checklists; (5) behaviorally anchored rating scales; (6) checklists; (7) computerized assessments; (8) the critical incident technique; (9) essays; (10) the in-basket test; (11) interviews; (12) logs or journals; (13) observations; (14) paper-and-pencil tests; (15) portfolios; (16) projects; (17) questionnaires; (18) rating forms; (19) track records; (20) video and audio tapes; (21) work sample tasks; and (22) work simulation tasks. To determine the method most appropriate for a given performance evaluation system, it is necessary to consider purpose and criteria, as well as individuals, resources, and legal and technical issues. (Contains 31 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 374 161

EREAPA ASSOCIATES

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
PATRICIA H. WHEELER
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

EREAPA Publication Series No. 93-6

Methods for Assessing Performance

Patricia H. Wheeler, M.B.A., Ph.D.

BEST COPY AVAILABLE

EREAPA
Publication
Series
No. 93-6

Methods for Assessing Performance

Patricia H. Wheeler, M.B.A., Ph.D.

EREAPA Associates
2840 Waverley Way
Livermore, California 94550-1740

1993

Copyright © 1993 by Patricia H. Wheeler.

The author expresses her appreciation to Geneva D. Haertel, Jean Martinson, and Paul K. Wheeler for their comments and suggested changes on earlier versions of this paper.

Methods for Assessing Performance

Patricia H. Wheeler, M.B.A., Ph.D.
EREAPA Associates
Livermore, California

A variety of assessment methods and instruments can and should be used to evaluate the performance of individuals and groups including students, learning groups, employees, applicants, teams, and program participants. Concerning the use of observations, portfolios, and other methods for the evaluation of teachers, Shulman (1988) points out,

> Each of these several approaches to the assessment of teachers is, in itself, as fundamentally flawed as it is reasonably suitable, as perilously insufficient as it is peculiarly fitting. What we need, therefore, is a union of insufficiencies, a marriage of complements, in which the flaws of individual approaches to assessment are offset by the virtues of their fellows. (p. 38)

It is important to use not only multiple sources of data such as clients, peers, and products (Wheeler, 1992b), but also a variety of assessment methods. When assessing teachers, for example, some domains and attributes are more appropriately measured by one or two of these methods, and other domains and attributes by different methods. The teacher's knowledge of the subject matter can be better assessed through portfolios, paper-and-pencil tests, and interviews rather than through observations. A classroom observation would allow for only a very limited sample for assessing this area, i.e., the topic being taught that day and questions on other topics asked by the students. If it is a pre-announced observation, then the teacher is likely to be teaching a topic in which he/she is quite knowledgeable and comfortable. Generally, students do not have a level of expertise in a subject area that would allow them to validly rate the teacher's knowledge of the subject matter. On the other hand, communication skills and management skills are better assessed through classroom observation and student ratings. Portfolios and questionnaires are probably most appropriate for such domains as the assessment of students and the teacher's service to the profession. Multiple sources of data, as well as multiple assessment methods, should be used for the evaluation of performance and decision-making about individuals.

Assessment Methods

Numerous possible assessment methods can be used to assess the performance of individuals and groups. Some are these are described below. They include what are called "alternative assessments" (i.e., assessment approaches that do not use multiple-choice or closed-response items). Many can be considered "authentic assessments," although the term "authentic" refers to the relationship of the assessment task to the performance being measured. As an example, performing cardio-pulmonary resuscitation (CPR) on a dummy is an authentic, real-life task for someone who wants to be a CPR instructor, since that person must be able to demonstrate CPR on a dummy to a group of students. For a first-aider or health professional, performing CPR on a dummy is an authentic, simulated task for performing CPR on a person. Completing a series of multiple-choice questions on CPR, or watching a video tape of someone performing CPR and listing the errors made, would be considered non-authentic assessments of ability to perform CPR. Informational resources on these assessment methods and others can be found in the "Bibliography."

Adaptive testing is an assessment process in which the sequence of items or tasks depends on responses to previous items or performance on previous tasks. Performance is assessed based on characteristics such as the difficulty of items answered correctly or types of tasks performed satisfactorily, rather than on how many items are answered correctly or how many tasks are completed satisfactorily. For instance, on a reading test if an individual responds correctly to a set of items on literal comprehension, then he/she will probably be given a set of items on inferential comprehension. However, if he/she is unable to answer most of the literal comprehension items, he/she may be given an easier set. A word processing person may be given a standard business letter to type. If done satisfactorily, then he/she may be given more complex materials to type, such as outlines, tables, and formulae. If he/she cannot do the letter, the next task might be a short paragraph or two.

An **anecdotal record** is a short narrative report or summary of an event or activity that may be used to document or support generalizations about the performance of an individual or a group, or about a situation. Examples include a memorandum for the record on how an employee dealt with a dissatisfied customer, or a short progress report that a project coordinator submitted to management.

Assessment center refers to the process of using simulation techniques to measure an individual's knowledge, performance, or potential. This term does not refer to a location, but instead describes an assessment approach that could be implemented in any of several locations. Several types of assessment exercises or tasks are used with an assessment center method. For example, potential plant managers might have to participate in a simulated interview for hiring a section supervisor, complete an in-basket test, analyze and compare financial reports for the past quarter and the last year, watch a video tape on a production process and make recommendations for improving it, and draft a memorandum to employees on a new policy.

A **behavioral checklist** is a record or form on which the assessor indicates the presence or absence of a specific behavior. The checklist may be designed to cover spontaneous and/or structured performance. For example, a doctor might notice that a child can tie his/her shoe (spontaneous) or the doctor might ask the child to tie his/her shoe (structured). An office manager might notice that an employee can operate certain equipment by observing that employee using the postal metering machine, sending a fax, changing the toner in the printer, and making collated copies (spontaneous behaviors). Or, the manager might ask the employee to perform each task at a designated time while the manager observes (structured).

A **behaviorally anchored rating scale**, commonly called BARS, is a type of assessment in which judgments about performance are empirically linked to specific examples of incumbent performance at each level of effectiveness on the scale. That is, the editor might rate the performance of a newspaper reporter on such attributes or domains as use of appropriate sources, adequate background knowledge before interviewing/writing, appropriate writing style, standard written and grammatically correct English, and so forth. For each of these, the assessor develops a scale of various performance levels, based on empirical data about performance of newspaper reporters. The editor compares the individual's performance to the various benchmark performance levels to determine what rating should be assigned. On the area of adequate background, performance levels might range from thorough familiarity to total ignorance concerning the issues and relevant background on the topic for a story.

A **checklist** is an instrument that specifies criteria or indicators of merit on which the assessor marks the presence or absence of the attribute being assessed. The criteria may or may not refer to specific behaviors, as with behavioral checklists (see above). A checklist may be used with college applicants to ensure that they meet certain requirements (e.g., graduation from an accredited high school, minimum high school grade-point average, minimum test scores, sufficient number of courses in each subject area, community service hours). The management of a bus company could use a checklist to assess a bus driver's performance by having the observer indicate if the driver followed the correct route, obeyed all traffic laws, made stops when passengers were present, and met the printed schedule within some reasonable deadline.

Computerized assessment uses computers to measure performance on some attribute, not necessarily related to computer skills. Computers can deliver multiple-choice tests and structured interview questions. Many instructional programs for young children include questions that are presented as each child progresses through the materials. Assessors can test pilots using computer simulations of landings. Computers can present accountants with financial data to manipulate and analytic problems to solve. An interview could be conducted with an applicant for a job or with a person completing a program. In these cases, the individuals enter responses to questions or situations shown on the computer screen. Computers present endless possibilities, especially with the era of virtual reality fast upon us.

The **critical incident technique**, developed in the mid-fifties by John Flanagan, uses documentation concerning critical incidents, that is, significant and observable episodes or performances (effective or ineffective) that alter the direction of subsequent events, behaviors, or activities. Such documentation may include information about: (1) noteworthy accomplishments (e.g., finishing a high quality report under a very tight deadline, landing a plane with one engine on fire, designing and implementing a cost-saving procedure that maintains product quality); (2) substantive improvement (e.g., learning how to use several new software packages, improving one's physical coordination through dance classes and exercise programs); and (3) significant failures (e.g., prescribing the wrong medication for a patient, overlooking extensive dry rot during a building inspection.)

An **essay** is an oral or written narrative description, analysis, explanation, interpretation, opinion, and/or summary that demonstrates an individual's use of information and language to generate a coherent document about a particular topic or in response to a question or prompt. Teachers often use essays to assess student learning and academic skills (e.g., book reports, essay questions on tests). Essays can also be used to assess such attributes as the motivation and career planning of college applicants, employees' understanding of the importance of following a certain policy, and program participants' reflections of their own success in changing behavior as a result of participating in a program.

The **in-basket test** is an assessment method wherein an individual receives a collection of documents, letters, memoranda, and other materials, each of which requires a decision or action. Within a specified time, the individual being assessed must review the materials, determine what actions need to be taken, and establish priorities for completing them. Some items that might be used in an in-basket test are: a letter of complaint from a client, a phone message about the failure to deliver materials on time to another site, a memo from a supervisor concerning inappropriate behavior of a long-term employee, a rush purchase request to order new capital equipment, a financial report on a project that is overspending in relation to its budget, and a request for a bid.

An **interview** consists of a series of verbally-delivered questions designed to elicit responses concerning attitudes, information, interests, knowledge, quick-response skills, and opinions. The interview can be in person, by telephone, or with small groups. The three types of interviews are structured, semi-structured, and unstructured. They differ in the degree of specificity of questions asked of the individual and in the extent to which the interviewer can use prompts and follow-up questions to obtain additional information or to clarify responses. In a structured interview, all questions to be asked are specified in advance and the interviewer cannot ask additional questions; such an interview could be delivered by computer or on a questionnaire with open-ended questions. Some questions are specified in advance for a semi-structured interview, but the interviewer can ask additional related or clarifying questions. For unstructured interviews, the interviewer uses a list of topics to develop questions and may also digress to questions on other topics.

A **log** is a journal or diary, maintained by an individual or group, that may include such topics as plans, activities, decisions, work undertaken and completed, time spent, products produced, results, changes, and reflections. The log can serve as a source of background, contextual, and performance information for use by a support provider, assessor, or evaluator. It can be included as part of a portfolio.

Observation is the careful watching and noting of behaviors and events. Observations typically occur in the individual's learning or work setting, but they also may occur in other settings (e.g., staff meetings, conferences, field trips) or may be based on audio or video tapes. Observation approaches include checklists, coding forms, frequency counts, rating forms, guided-note taking records, and scripting. Observations may be pre-announced or unannounced. They can focus on certain individuals, or aspects of performance, or be broad-based in terms of what behaviors and events are covered.

A **paper-and-pencil test** consists of items, questions, or problems to be answered by the individual in writing or by marking an answer document (e.g., checking a box, filling in a bubble). Usually these tests consist of multiple-choice items (e.g., matching, true-false), but may also include other types of items (e.g., fill in the blank, underline the error, label a diagram or map). Most often they are used in academic settings and with large-scale testing programs because, once developed, they usually are less costly to administer and score than other forms of assessment (Wheeler, 1992a).

A **portfolio** is a purposeful collection of: (1) documents concerning an individual's performance (e.g., assessment results, awards, testimonials from clients, peer-evaluations, supervisor reports, summary of training completed), and (2) products produced by the individual (e.g., reports, action research results, self-evaluations, reflective essays, video tapes of work activities, audio tapes of talks given to groups or conferences with clients, photographs of work accomplishments).

A **project** is a form of complex performance assessment involving several types of activities for completion within a specified period of time. Most projects start with planning and lead to a product such as an oral or written report, or a display or scale model. Projects can be done by individuals or groups. Examples of projects are: completing several data collection activities on local transportation patterns and preparing a set of recommendations for the city council; planning an experiment on soil erosion and setting up various situations to illustrate what happens under each condition and to compare results; designing a piece of furniture and building a prototype; doing library

research and interviews on a local history topic and making a presentation in class; or writing a doctoral dissertation or thesis.

A **questionnaire** consists of a series of queries and statements, and is used to collect data and information. An individual can response to questions on such topics as educational background, career plans, training completed, work experience, special skills, attitudes, opinions, reactions, and community service activities. Or others (e.g., clients, peers, supervisors, students) can complete a questionnaire concerning an individual's performance. An application form is a type of questionnaire.

A **rating form** is an instrument on which the magnitude or degree of some aspect of performance is estimated. Such forms may use a numerical continuum (e.g., 1-2-3-4) or a descriptive continuum (e.g., well-good-fair-poor; frequently-fairly often-sometimes-rarely-never). Ratings are usually based on scale definitions, scoring rubrics, and benchmarks, not in relation to the performance of other individuals being assessed, as with rankings. A commonly used format for rating scales is the Likert scale in which individuals are given a continuum of responses from which to select (e.g., strongly agree to strongly disagree, frequently to never, very helpful to not helpful).

A **track record** is a summary of past events and accomplishments related to an individual's performance. In addition to information about past performance, a track record may include details of further education and training completed, conferences and meetings attended, awards received, disciplinary actions, attendance records, products produced, and noteworthy accomplishments. Such information is usually included in the individual's portfolio and on the résumé or vita.

Video and audio tapes are recordings of an individual performing such tasks as implementing a procedure, participating in an activity with others, conferring with supervisors or other staff, and interacting with clients. An assessor reviews the tapes and makes notes about behaviors and events, in much the same manner as an observation. Tapes permit the assessor to review behaviors and events several times to ensure that notes are accurate and complete. However, tapes can miss much of what is happening, since they only record a limited amount of information. One cannot see what is happening in areas not covered by the camera, nor can one see gestures, audience reactions, and visual aides if there is only an audio tape.

A **work sample task** is an actual or typical activity used to assess performance (e.g., asking a librarian to locate certain information in the reference section, having a translator interview someone and provide a written report in another language).

A **work simulation task** is a surrogate or imitation of a sample task that is used to assess performance (e.g., asking a librarian where one should look for certain information in the reference section, having a translator listen to a tape and answer questions about the content of what was said).

Issues to Address

To determine which assessment methods are most appropriate for a given performance evaluation system, several key policy decisions must be addressed as part of the selection or development of assessment methods and instruments. These include:

(1) the purpose of the assessment and evaluation system;

(2) the criteria to be covered by the assessment system, and the domains and indicators associated with those criteria;

(3) the individuals who will collect the assessment data or administer the assessment instruments, including the individuals being evaluated, managers, supervisors, peers, mentors or support providers, and clients;

(4) the individuals who will use the assessment data, including the individuals being evaluated, managers, supervisors, evaluators, mentors or support providers, and clients;

(5) resources available for the assessment and evaluation (e.g., people, time, equipment, rooms, materials);

(6) technical issues including reliability, accuracy, relevance to job responsibilities or program goals, fairness and objectivity, validity, and comparability across settings and assignments; and

(7) legal issues including authenticity, appeals procedures, compliance with union agreements, confidentiality of information and materials, and protection against misuse of the data or procedures associated with the assessment.

In summary, two primary reasons for using multiple assessment methods are: (1) no one instrument or method is appropriate for all aspects of performance covered by the evaluation system; and (2) the use of multiple assessment methods for one domain or attribute of performance allows the verification of data and triangulation of results.

Bibliography

These items may be useful to those seeking further information on assessment methods. References cited in this paper are included in this bibliography.

Angelo, Thomas A.; & Cross, K. Patricia. (1993). *Classroom assessment techniques: A handbook for college teachers*. San Francisco, CA: Jossey-Bass Inc., Publishers.

Arter, Judith A.; & Spandel, Vicki. (1992, Spring). Using portfolios of student work in instruction and assessment. *Educational Measurement: Issues and Practice*, *11*(1), 36-44.

Berk, Ronald A. (Ed.). (1986). *Performance assessment: Methods and applications*. Baltimore, MD: The Johns Hopkins University Press.

Brookhart, Susan. (1993, Spring). Assessing student achievement with term papers and written reports. *Educational Measurement: Issues and Practice*, *12*(1), 40-47.

Eder, Robert W.; & Ferris, Gerald R. (Eds.). (1989). *The employment interview: Theory, research and practice*. Newbury Park, CA: Sage Publications, Inc.

Flanagan, John C. (1954). The critical incident technique. *Psychological Bulletin*, *51*, 327-358.

Haertel, Edward H. (1986, Spring). The valid use of student performance measures for teacher evaluation. *Educational Evaluation and Policy Analysis*, *8*(1), 45-60. (ERIC Document Reproduction Service No. EJ 350 184)

Haertel, Edward H. (1991). New forms of teacher assessment. In Gerald Grant (Ed.), *Review of research in education* (Vol. 17) (pp. 3-29). Washington, DC: American Educational Research Association.

Herman, Joan L.; Aschbacher, Pamela R.; & Winters, Lynn. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hirabayashi, Judy; & Wheeler, Patricia. (1992). *Approaches to classroom observations: Open versus closed systems* (EREAPA Publication Series No. 92-8). Livermore, CA: EREAPA Associates.

Jacobs, Lucy Cheser; & Chase, Clinton I. (1992). *Developing and using tests effectively: A guide for faculty*. San Francisco, CA: Jossey-Bass Inc., Publishers.

Landy, Frank J.; & Farr, James L. (1980). Performance ratings. *Psychological Bulletin*, *87*, 72-107.

Millman, Jason; & Darling-Hammond, Linda. (Eds.). (1990). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, CA: Sage Publications, Inc.

Morris, Lynn Lyons; Fitz-Gibbons, Carol Taylor; & Lindheim, Elaine. (1987). *How to measure performance and use tests*. Newbury Park, CA: Sage Publications, Inc.

Paulson, F. Leon; Paulson, Pearl; & Meyer, Carol. (1991). What makes a portfolio a portfolio? *Educational Leadership*, *48*(5), 60-63.

Priestly, M. (1982). *Performance assessment in education and training: Alternative techniques*. Englewood Cliffs, NJ: Educational Technology Publications.

Schwab, Donald P.; Heneman, Herbert G , III; & DeCotis, Thomas A. (1975). Behaviorally anchored rating scales: A review of the literature. *Personnel Psychology, 28*, 549-562.

Scriven, Michael. (1987). *The validity of student ratings*. Unpublished manuscript.

Scriven, Michael. (1991). *Evaluation thesaurus* (4th ed.). Newbury Park, CA: Sage Publications, Inc.

Shaw, M. E.; & Wright, J. M. (1967). *Scales for the measurement of attitudes*. New York, NY: McGraw-Hill.

Shulman, Lee S. (1988, November). A union of insufficiencies: Strategies for teacher assessment in a period of educational reform. *Educational Leadership, 46*(3), 36-41. (ERIC Document Reproduction Service No. EJ 385 344)

Smith, P.; & Kenfall, L. M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Stiggins, Richard. (1987, Fall). Design and development of performance assessments. *Educational Measurement: Issues and Practice, 6*(3), 33-42.

Thornton, George C., III; & Byham, William C. (Eds.). (1982). *Assessment centers and managerial performance*. New York, NY: Academic Press.

Walberg, Herbert J.; & Haertel, Geneva D. (Eds.). (1991). *The international encyclopedia of educational evaluation*. Oxford, England: Pergamon Press.

Wheeler, Patricia. (1992a). *Relative costs of various types of assessments* (EREAPA Publication Series No. 92-2). Livermore, CA: EREAPA Associates.

Wheeler, Patricia. (1992b, October). *Sources of data for evaluating teachers* (TEMP Memo 7). Kalamazoo, MI: Western Michigan University, The Evaluation Center, Center for Research on Educational Accountability and Teacher Evaluation.

Wheeler, Patricia. (1993). *Using portfolios to assess teachers* (EREAPA Publication Series No. 93-7). Livermore, CA: EREAPA Associates.

Wheeler, Patricia; & Haertel, Geneva D. (1993). *Resource handbook on performance assessment and measurement: A tool for students, practitioners, and policymakers*. Berkeley, CA: The Owl Press.

Wheeler, Patricia; Haertel, Geneva D.; & Scriven, Michael. (1992). *Teacher evaluation glossary*. Kalamazoo, MI: Western Michigan University, The Evaluation Center, Center for Research on Educational Accountability and Teacher Evaluation.

Wolf, Kenneth. (1991, October). The schoolteacher's portfolio: Issues in design, implementation, and evaluation. *Phi Delta Kappan, 73*(2), 129-136.